Dual-Directed Algorithm Design for Efficient Pure Exploration

Wei You[1]

香港科技大学

[1]Joint work with Chao Qin (Stanford University). To appear in *Operations Research*. Draft available at https://arxiv.org/abs/2310.19319.

Outline

Adaptive experiments

In stochastic adaptive experiments, the decision maker

- faces a finite set of alternative options;
- the mean performance θ_i is **unknown**, whose uncertainty can only be reduced by costly experiments or measurements.

We seek to allocate measurement efforts wisely to correctly answer a query about the alternatives with high confidence.

Exploration query

An *exploration query* specifies a question to be answered regarding the unknown mean vector $\boldsymbol{\theta}$.

Examples of pure exploration queries

We study pure-exploration problems in adaptive experiments with a finite set of candidates.

1. One may seek to quickly identify *the best performing alternative* $I^* = \text{argmax}_i \theta_i$.
   - a.k.a. "*best arm identification*" or "*ranking and selection*".
   - **Applications**: Hyperparameter tuning[1], LLM prompt optimization[2], brain-computer interface[3].
   - Variants of this problem
     - Finding one or all good enough alternatives, i.e., i such that $\theta_i > \theta_{I^*} - \varepsilon$.
     - Finding the best-k alternatives.
     - Finding a subset that contains the best alternative.

[1] X. Shang, E. Kaufmann, and M. Valko, "A simple dynamic bandit algorithm for hyper-parameter tuning," in *6th ICML Workshop on Automated Machine Learning*, 2019.

[2] R. Pryzant, D. Iter, J. Li, Y. T. Lee, C. Zhu, and M. Zeng, "Automatic Prompt Optimization with ``Gradient Descent'' and Beam Search," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[3] X. Zhou, B. Hao, T. Lattimore, J. Kang, and L. Li, "Sequential Best-Arm Identification with Application to P300 Speller," *Transactions on Machine Learning Research*, 2024.

Examples of pure exploration queries

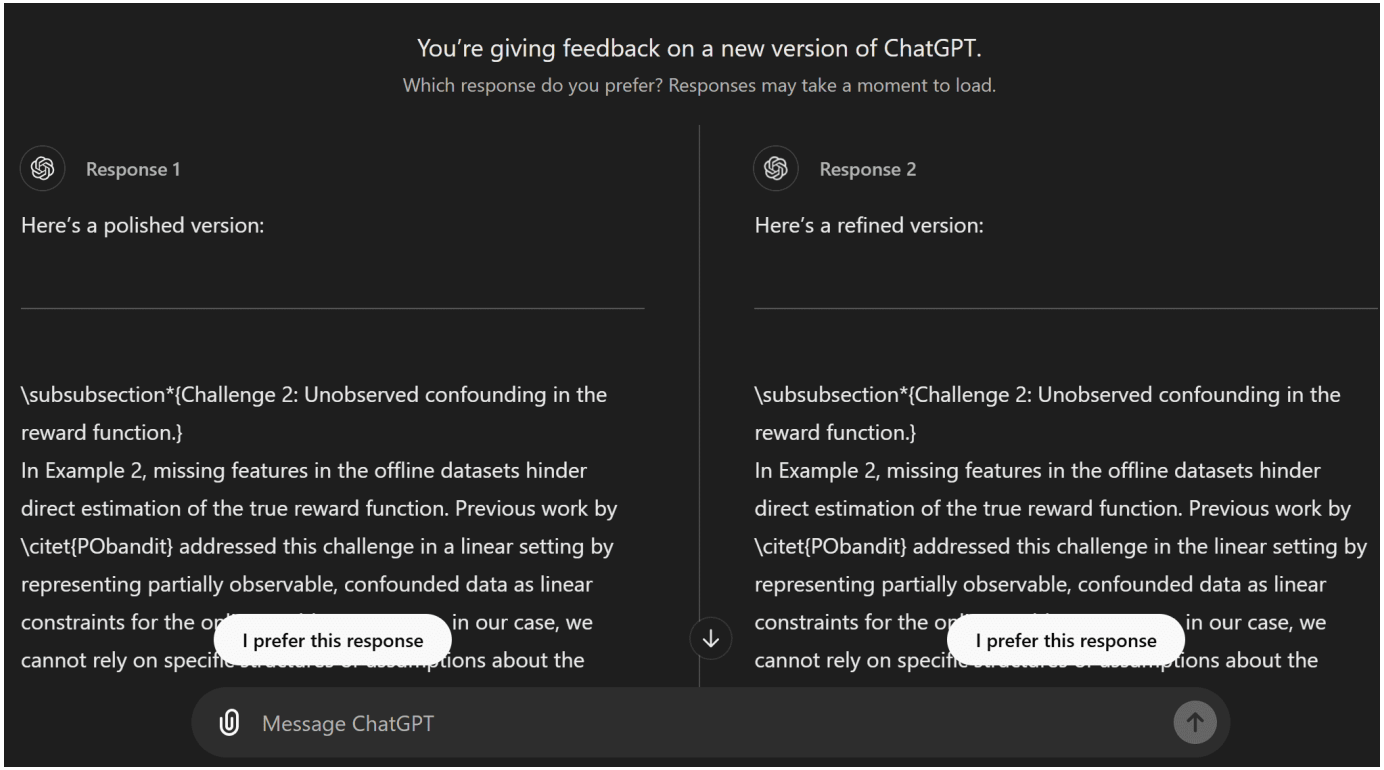
2. One may seek to compare the mean performance against some threshold
   - Find all alternatives above a threshold, a.k.a. *thresholding bandits*[1].
   - Variants of this problem
     - Find the alternatives closest to the threshold.
     - Verify if the smallest mean is lower than a threshold, a.k.a. *Murphy sampling*[2].

[1] A. Locatelli, M. Gutzeit, and A. Carpentier, "An optimal algorithm for the thresholding bandit problem," in *International Conference on Machine Learning*, 2016, pp. 1690–1698.

[2] E. Kaufmann, W. M. Koolen, and A. Garivier, "Sequential test for the lowest mean: From Thompson to Murphy sampling," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

Examples of pure exploration queries

3. Weeks before the official release of **`OpenAI o3`** and **`o4-mini`** on April 16, 2025, they extensively tested versions of the LLM using pure exploration in dueling bandits[1].



[1] V. Dwaracherla, S. M. Asghari, B. Hao, and B. Van Roy, "Efficient exploration for LLMs," *arXiv preprint arXiv:2402.00396*, 2024.

What is pure exploration?

We seek to answer a query about the unknown mean vector $\boldsymbol{\theta}$ with high confidence, while *using as few measurements as possible*. The emphasis is on *end outcomes*.

- Costs incurred during the adaptive experimentation phase is high and does not depend on the mean performance.
- Incur potentially large costs after the experiment.
  - Long-term commitment of resources based on the answer to the query.
  - **Example**:
    - Mass production of a product.
    - Tenure promotion.
    - Deploy a LLM.
    - Construction of a new hospital.

Pure exploration

Pure exploration

Exploration matters the most, whereas exploitation is unnecessary.

An **inherent trade-off**:

- Minimizing the length of the experimentation phase.
- Answering the query correctly with high confidence.

Various formulations of pure exploration:

- **Fixed-budget**: Maximize accuracy under a fixed number of measurements[1].
- **Fixed-confidence**: Minimize the number of measurements to guarantee a given accuracy.

[1] An alternative is the posterior fixed-budget setting: minimize the large deviation rate of the posterior probability of an incorrect answer, under a fixed number of measurements.

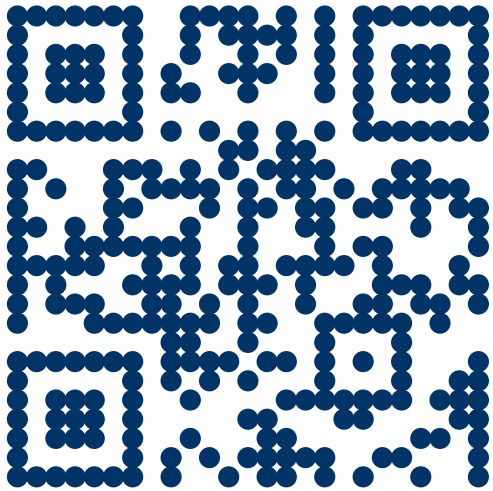
Motivations

We are motivated by the need of **a unified algorithm design principle** for

- different *problem formulations* (fixed-budget, fixed-confidence, etc.).
- different *exploration queries* (best arm identification, thresholding, etc.).
- different *noise distributions* (Gaussian, Bernoulli, etc.).

We are particularly interested in *variants of* ***Thompson sampling (TS)*** for this purpose.

Access our full paper here:



Outline

Thompson sampling

In bandit literature, a popular formulation is *regret minimization*.

Regret minimization The goal is to minimize the expected regret, defined as the difference between the expected reward of the optimal arm and the expected reward of the chosen arm.

Thompson sampling (TS) is a popular algorithm for regret minimization.

Thompson sampling[1]

- After each observation, update the posterior distribution Π_t of the mean vector θ.
- Draw a sample $\tilde{\theta}$ from Π_t.
- Choose the arm with the largest sample: $I_t = \operatorname{argmax}_{i \in [K]} \tilde{\theta}_i$.

[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

Can we apply TS to best-arm identification?

TS is designed for regret minimization.

- In minimizing the regret, TS must allocate sufficient measurements to the best-performing arms (exploitation), however, these exploitation efforts does not contribute to the goal of pure exploration.

Example: Consider two normal populations $\mathcal{N}(\theta_1, 1)$ and $\mathcal{N}(\theta_2, 1)$ with $\theta_1 > \theta_2$. We test

$$H_0 : \theta_1 \leq \theta_2 \quad \text{versus} \quad H_1 : \theta_1 > \theta_2.$$

Consider the Z-statistics $Z_t \stackrel{\text{def}}{=} \frac{\overline{X}_t - \overline{Y}_t}{\sqrt{\frac{1}{n_{1,t}} + \frac{1}{n_{2,t}}}}$.

- The larger Z_t is, the more likely we reject H_0.
- TS has $Z_t = O(\log(t))$. But equal allocation result in $Z_t = O(t)$.

A simple modification of TS for BAI

Thompson sampling

- Update posterior Π_t.
- Sample $\tilde{\boldsymbol{\theta}}$ from Π_t.
- Choose greedily $I_t = \operatorname{argmax}_{i \in [K]} \tilde{\theta}_i$.

When the tuning parameter β is set to 1, top-two TS is equivalent to TS.

The name *top-two* refers to the two candidates, leader and challenger.

Top-two Thompson sampling[1]

- Update posterior Π_t.
- **Leader**: Sample $\tilde{\boldsymbol{\theta}}$ from Π_t and let $I_t^{(1)} = \operatorname{argmax}_{i \in [K]} \tilde{\theta}_i$.
- **Challenger**: Sample repeatedly from Π_t until $\tilde{\boldsymbol{\theta}}'$ such that $I_t^{(2)} = \operatorname{argmax}_{i \in [K]} \tilde{\theta}'_i \neq I_t^{(1)}$.
- **Tuning**: play the leader with probability β and the challenger with probability $1 - \beta$.

[1] D. Russo, "Simple Bayesian Algorithms for Best-Arm Identification," *Operations Research*, vol. 68, no. 6, pp. 1625–1647, 2020.

Research questions and contributions

1. How to remove the tuning parameter β? Open problem[1].
   - We provide a *parameter-free algorithm.*
2. How to modify TS to solve other pure exploration problems?
   - We provide surprisingly simple variants of TS for other pure-exploration problems.
3. Is the parameter-free algorithm optimal?
   - Yes, optimality is established for Gaussian BAI.
4. How to obtain computationally efficient TS variants?
   - As posterior concentrates, the repeated sampling step takes a long time.
   - We provide two computationally efficient variants that are optimal.

[1] D. Russo, "Simple Bayesian Algorithms for Best-Arm Identification," *Operations Research*, vol. 68, no. 6, pp. 1625–1647, 2020.

Outline

Learning objective

Correct answer

The pure-exploration query induces a correct answer $\mathcal{I}(\boldsymbol{\theta})$, which we assume is unique.

Fixed-confidence performance criteria Given a confidence level $\delta \in (0,1)$. At each time,

- **Stopping rule**: the DM check if a stopping condition is met.
- **Decision rule**: If met, the DM stops and outputs the answer $\hat{\mathcal{I}}_{\tau_\delta}$.
- **Selection rule**: Otherwise, the DM selects an arm I_t.

δ-correct

A policy is said to be δ-correct if the *probability of correct selection* (PCS) upon stopping at time τ_δ is at least $1-\delta$, i.e.,

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\tau_\delta < \infty, \hat{\mathcal{I}}_{\tau_\delta} = \mathcal{I}(\boldsymbol{\theta})\right) \geq 1-\delta, \quad \text{for all } \boldsymbol{\theta}.$$

Learning objective

We seek to find the policy that is δ-*correct*, while *minimizing the expected stopping time*.

Universal efficiency

A policy π^* is said to be *universally efficient* if π^* is δ-correct and for any other δ-correct policy π, we have

$$\frac{\mathbb{E}_{\boldsymbol{\theta}}^{\pi^*}[\tau_\delta]}{\mathbb{E}_{\boldsymbol{\theta}}^{\pi}[\tau_\delta]} \leq 1, \quad \text{for all } \boldsymbol{\theta}.$$

Decomposition of pure exploration tasks

Observation

A pure exploration task can often be decomposed into *simpler tasks*.

Example (Best-arm identification):

We test if arm I^* is indeed the true best arm. It is equivalent to testing multiple two-arm hypotheses:

$$H_{0,x} : \theta_x > \theta_{I^*} \quad \text{versus} \quad H_{1,x} : \theta_x \leq \theta_{I^*}, \quad \text{for all } x \in [K] \setminus \{I^*\}.$$

- Each pair of hypotheses checks if a sub-optimal arm x is better than I^*.
- If we fail to reject $H_{0,x}$, then x is the reason that I^* does not appear as the best arm under the data. Hence, we refer to x as a possible *pitfall*.

Decomposition of pure exploration tasks

Let $\mathcal{I}(\boldsymbol{\theta})$ denote the unique correct answer.

- For an algorithm to correctly answer the pure-exploration query, it must distinguish between problem instances that yield different answers.

We collect all alternative parameters that leads to a different answer as the *alternative set*.

Pitfalls and decomposition of alternative set

We assume that $\mathrm{Alt}(\boldsymbol{\theta})$ can be decomposed into the union of a finite set of convex sets:

$$\mathrm{Alt}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \{\boldsymbol{\vartheta} : \mathcal{I}(\boldsymbol{\vartheta}) \neq \mathcal{I}(\boldsymbol{\theta})\} = \cup_{x \in \mathcal{X}} \mathrm{Alt}_x(\boldsymbol{\theta}).$$

We refer to $\mathcal{X}$ the set of *pitfalls*.

Example (Best-arm identification): $\mathcal{X} = [K] \setminus \{I^*\}$, and $\mathrm{Alt}_x(\boldsymbol{\theta}) = \left\{\boldsymbol{\vartheta} : \vartheta_x > \vartheta^*_{\hat{I}}\right\}$

A hypothesis testing perspective for general pure exploration

We test if arm $\mathcal{I}$ is indeed the correct answer. It is equivalent to testing multiple hypotheses:

$$H_{0,x} : \boldsymbol{\theta} \in \mathrm{Alt}_x(\boldsymbol{\theta}) \quad \text{versus} \quad H_{1,x} : \boldsymbol{\theta} \notin \mathrm{Alt}_x(\boldsymbol{\theta}), \quad \text{for all } x \in \mathcal{X}.$$

The generalized log-likelihood ratio test (GLRT) statistic is given by

$$\ln \frac{\sup_{\boldsymbol{\vartheta}} L(\boldsymbol{\vartheta})}{\sup_{\boldsymbol{\vartheta} \in \mathrm{Alt}_x(\boldsymbol{\theta})} L(\boldsymbol{\vartheta})} = t \cdot \inf_{\boldsymbol{\vartheta} \in \mathrm{Alt}_x(\boldsymbol{\theta})} \sum_{i \in [K]} p_i d(\theta_i, \vartheta_i),$$

where p_i is the proportion of samples allocated to arm i and $d(\cdot, \cdot)$ is the Kullback-Leibler (KL) divergence.

Generalized Chernoff information (information gain per sample)

$$C_x(\boldsymbol{p}; \boldsymbol{\theta}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{\vartheta} \in \mathrm{Alt}_x(\boldsymbol{\theta})} \sum_{i \in [K]} p_i d(\theta_i, \vartheta_i), \quad \text{for all } x \in \mathcal{X}.$$

Problem complexity

Recall that

- All $H_{0,x} : \boldsymbol{\theta} \in \mathrm{Alt}_x(\boldsymbol{\theta})$ must be rejected simultaneously to declare $\mathcal{I}$ the answer.
- $C_x(\boldsymbol{p}; \boldsymbol{\theta})$ quantifies the information gathered to reject $H_{0,x}$.

The DM seek to solve

$$\Gamma_{\boldsymbol{\theta}}^* \stackrel{\text{def}}{=} \max_{\boldsymbol{p} \in \mathcal{S}_K} \min_{x \in \mathcal{X}} C_x(\boldsymbol{p}; \boldsymbol{\theta}), \quad \boldsymbol{p}^* \in \operatorname*{argmax}_{\boldsymbol{p} \in \mathcal{S}_K} \min_{x \in \mathcal{X}} C_x(\boldsymbol{p}; \boldsymbol{\theta}),^1$$

where $\mathcal{S}_K \subset \mathbb{R}^K$ denote the probability simplex.

💡 Idea

Wisely allocate the measurement efforts to maximize the smallest test statistic.

[1] ⚠️ For general problems, it is possible that the optimal solution is non-unique. Is this especially true for bandit with structures, such as linear bandits.

A sufficient condition for optimality

Theorem (Y. and Qin, 2024[1])

Using appropriate stopping rule, if a algorithm ensures that

$$\boldsymbol{p}_t \overset{\mathbb{M}}{\to} \boldsymbol{p}^*,$$

then it is universally efficient. In particular,

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\theta}}[\tau_\delta]}{\log(1/\delta)} = \frac{1}{\Gamma^*_{\boldsymbol{\theta}}}.$$

- Matching lower bound established in[2][3].
- The problem is then reduced to finding sampling rules that rapidly converges to $\boldsymbol{p}^*$.

---

[1] C. Qin and W. You, "Dual-directed algorithm design for efficient pure exploration," *arXiv preprint arXiv:2310.19319*, 2024.

[2] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Conference on Learning Theory*, 2016, pp. 998–1027.

[3] P.-A. Wang, R.-C. Tzeng, and A. Proutiere, "Fast Pure Exploration via Frank-Wolfe," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

The optimal allocation problem

$$\Gamma_{\boldsymbol{\theta}}^* = \max_{\boldsymbol{p} \in \mathcal{S}_K} \min_{x \in \mathcal{X}} C_x(\boldsymbol{p}; \boldsymbol{\theta})$$

Theorem (Optimality conditions, Y. and Qin, 2024[1])

$$[\text{Stationarity}] \quad p_i = \sum_{x \in \mathcal{X}} \mu_x h_i^x(\boldsymbol{p}), \quad \forall i \in [K],$$

$$[\text{Complementary slackness}] \quad \mu_x \left(\min_{x' \in \mathcal{X}} C_{x'}(\boldsymbol{p}; \boldsymbol{\theta}) - C_x(\boldsymbol{p}; \boldsymbol{\theta}) \right) = 0, \quad \forall x \in \mathcal{X},$$

where μ_x is the **dual variable** for the inner minimization problem, and

$$h_i^x(\boldsymbol{p}) = \frac{p_i \frac{\partial C_x(\boldsymbol{p})}{\partial p_i}}{C_x(\boldsymbol{p})}.$$

[1] C. Qin and W. You, "Dual-directed algorithm design for efficient pure exploration," *arXiv preprint arXiv:2310.19319*, 2024.

Outline

The first design principle: the complementary slackness condition

Now we are ready to present our algorithm design principle.

First, consider the complementary slackness condition

$$\mu_x \left(\min_{x' \in \mathcal{X}} C_{x'}(\boldsymbol{p}; \boldsymbol{\theta}) - C_x(\boldsymbol{p}; \boldsymbol{\theta}) \right) = 0, \quad \forall x \in \mathcal{X}.$$

- **Interpretation for the dual variables μ_x**: the proportion of iterations where x is identified as the hardest alternative hypothesis (i.e., the principal pitfall).

💡 The first algorithm design principle (asymptotic version)

We identify x as the *principal pitfall* if

$$x = \operatorname*{argmin}_{x' \in \mathcal{X}} C_{x'}(\boldsymbol{p}; \boldsymbol{\theta}).$$

Connection to top-two TS

TTTS for BAI

Challenger: Sample repeatedly from Π_t until $\tilde{\boldsymbol{\theta}}'$ such that $I_t^{(2)} = \operatorname{argmax}_{i \in [K]} \tilde{\theta}'_i \neq I_t^{(1)}$.

Proposition 5[1]

TTTS asymptotically samples the arm with the smallest $C_x(\boldsymbol{p}; \boldsymbol{\theta})$ as challenger.

This observation provides the foundation of our modification to TS.

💡 The first algorithm design principle (Thompson sampling version)

Repeatedly sample from the posterior distribution untial an alternative answer emerges.

[1] D. Russo, "Simple Bayesian Algorithms for Best-Arm Identification," *Operations Research*, vol. 68, no. 6, pp. 1625–1647, 2020.

Examples of TS variants for pure exploration

Example (Best-k identification)

We wish to find the exact set of k arms whose mean is the best-k.

- Suppose the current sample mean tell us that the empirical best-k arms are $\mathcal{I}\left(\hat{\boldsymbol{\theta}}\right) = \{1, 2, ..., k\}$. Then we repeatedly sample $\tilde{\boldsymbol{\theta}}$ from the posterior distribution until $\mathcal{I}\left(\tilde{\boldsymbol{\theta}}\right) \neq \{1, 2, ..., k\}$, e.g., $\mathcal{I}\left(\tilde{\boldsymbol{\theta}}\right) = \{1, 2, ..., k-1, k+1\}$.
- **TS variant**: The reason that $\mathcal{I}\left(\tilde{\boldsymbol{\theta}}\right)$ leads to a different answer is because the order of arm k is swapped with $k+1$. Hence, the pair $(k, k+1)$ is the principal pitfall.
- We should collect more samples to further compare arm k with arm $k+1$.
  - **Candidates:** Notice that only new samples to arms k and $k+1$ contributed to a better understanding of the comparison between these two arms.

Examples of TS variants for pure exploration

Example (All ε-good arms)

We wish to identify all arms that are ε-good, i.e., i such that $\theta_i > \theta_{I^*} - \varepsilon$.

- **TS variant**: Suppose the current empirical answer is $\mathcal{I}\left(\hat{\boldsymbol{\theta}}\right) = \{1, 2\}$, then we repeadedly sample $\tilde{\boldsymbol{\theta}}$ until we find an arm $j \neq 1, 2$ such that $\tilde{\theta}_j > \tilde{\theta}_i + \varepsilon$ for $i = 1$ or 2.
- The reason that $\mathcal{I}\left(\tilde{\boldsymbol{\theta}}\right)$ leads to a different answer is because arm j certifies that arm i cannot be ε-good.
- We should collect more samples to further compare arm i with arm j.
  - **Candidates:** Notice that only new samples to arms i and j contributed to a better understanding of the comparison between these two arms.

Examples of TS variants for pure exploration

Example (Subset selection)

We wish to identify a subset of cardinality k that contains the best arm I^*.

- **TS variant**: Note that the best guess of the subset is the set with the k arms with the highest sample means. Suppose the set is $\mathcal{I}\left(\hat{\boldsymbol{\theta}}\right) = \{1, 2, \ldots, k\}$, then we repeatedly sample $\tilde{\boldsymbol{\theta}}$ from the posterior distribution until an arm j emerges such that $\tilde{\theta}_j > \max_{i \in [k]} \tilde{\theta}_i$.
- The reason that the existence of j leads to a different answer is because arm j certifies that none of the arms $i \in [k]$ is the best.
- We should collect more samples to further compare j with $i^* = \operatorname{argmax}_{i \in [k]} \hat{\theta}_i$. This is because i^* is the arm most indistinguishable from j.
  - **Candidates:** Notice that only new samples to arms i^* and j contributed to a better understanding of the comparison between these two arms.

How to choose from the candidate set?

In all previous examples, we

- first identify a principal pitfall x by TS;
- then the pitfall x identifies a set of **candidates** whose additional sample will contribute to the its mitigation.

Question: how to choose from the set of candidates to maximize information gain?

How to choose from the candidate set?

It turns out that the active candidate set $\mathfrak{C}_x$ for general pure exploration takes a simple form.

Active candidate set $\mathfrak{C}_x$

$$\mathfrak{C}_x = \left\{ i \in [K] : \frac{\partial C_x(\boldsymbol{p})}{\partial p_i} > 0 \right\}.$$

- Intuitively, only the arms that contribute to the information gain of $C_x(\boldsymbol{p}; \boldsymbol{\theta})$ are active.

The selection function

The selection function

$h_i^x(\boldsymbol{p})$ is called the *selection function* under pitfall x:

$$h_i^x(\boldsymbol{p}) = \frac{p_i \frac{\partial C_x(\boldsymbol{p})}{\partial p_i}}{C_x(\boldsymbol{p})}, \quad \text{where } \operatorname{Supp}(\boldsymbol{h}^x) = \mathfrak{C}_x.$$

- Intuitively, $h_i^x(\boldsymbol{p})$ is the proportion of information contributed by samples allocated to arm i for testing $H_{0,x} : \boldsymbol{\theta} \in \operatorname{Alt}_x(\boldsymbol{\theta})$.
- It can be verified that $\boldsymbol{h}^x(\boldsymbol{p}) = (h_1^x(\boldsymbol{p}), \ldots, h_K^x(\boldsymbol{p}))$ is a probability vector.

The second design principle: the stationarity condition

Recall that the stationarity condition is given by

$$p_i = \sum_{x \in \mathcal{X}} \mu_x h_i^x(\boldsymbol{p}), \quad \forall i \in [K].$$

- **The dual variables μ_x**: the proportion of times that x is identified as the principal pitfall.
- **The selection function $h_i^x(p)$**: the probability of selecting arm i under pitfall x.
- The stationarity condition is essentially the **law of total probability**.

💡 The second algorithm design principle (information-directed selection, IDS)

For a given pitfall x, select the arm $i \in \mathfrak{C}_x$ with probability $h_i^x(\boldsymbol{p})$.

Discussion

How about greedy rule? This was suggeted by[1], i.e., $i \in \operatorname{argmax}_{x \in \mathcal{X}} h_i^x(\boldsymbol{p})$.

- We have an example showing that greedy rule is not optimal.
- In essence, allocation must consider the *long-term average effect* of different x showing up as the principal pitfall, and react according to this average (i.e., the dual variable $\boldsymbol{\mu}$).

[1] P. Ménard, "Gradient Ascent for Active Exploration in Bandit Problems." [Online]. Available: https://arxiv.org/abs/1905.08165

The proposed algorithms

Proposed allocation rule (asymptotic)

- **[Estimate]**: Calculate estimate $\boldsymbol{\theta}_t$ of $\boldsymbol{\theta}$. (Default: sample mean.)
- **[Detect]**: detect the principal pitfall

$$x_t = \operatorname*{argmin}_{x' \in \mathcal{X}(\boldsymbol{\theta}_t)} C_{x'}(\boldsymbol{p}_t; \boldsymbol{\theta}_t).$$

- **[Select]**: Draw an arm I_t from the distribution $\boldsymbol{h}^{x_t}(\boldsymbol{p}_t)$.
- Pull arm, observe reward, update history, and advance time.

Proposed allocation rule (TS version)

- **[Estimate]**: Calculate estimate $\tilde{\theta}$ of θ. (Default: posterior sample.)
- **[Detect]**: Repeatedly sample $\tilde{\boldsymbol{\theta}}' \sim \Pi_t$ until $\tilde{\boldsymbol{\theta}}' \in \mathrm{Alt}\left(\tilde{\boldsymbol{\theta}}\right)$. Detect pitfall

$$x_t \in \left\{x \in \mathcal{X}\left(\tilde{\boldsymbol{\theta}}\right) : \tilde{\boldsymbol{\theta}}' \in \mathrm{Alt}_x\left(\tilde{\boldsymbol{\theta}}\right)\right\},$$

breaking tie arbitrarily.

- **[Select]**: Draw an arm I_t from the distribution $\boldsymbol{h}^{x_t}(\boldsymbol{p}_t)$.
- Pull arm, observe reward, update history, and advance time.

Optimality

Theorem (Y. and Qin, 2024[1])

For Gaussian BAI, ε-BAI, and thresholding bandits, our algorithm is universally efficient.

[1] C. Qin and W. You, "Dual-directed algorithm design for efficient pure exploration," *arXiv preprint arXiv:2310.19319*, 2024.

Summary and perspectives

The core ingredients of our algorithm are:

The maximin characterization of the optimal allocation problem

$$\Gamma_{\boldsymbol{\theta}}^* = \max_{\boldsymbol{p} \in \mathcal{S}_K} \min_{x \in \mathcal{X}} C_x(\boldsymbol{p}; \boldsymbol{\theta}).$$

- This maximin structure induces **the first design principle**.

The information decomposition

$$C_x(\boldsymbol{p}; \boldsymbol{\theta}) = \sum_{i \in [K]} p_i \frac{\partial C_x(\boldsymbol{p})}{\partial p_i}, \quad \text{for all } x \in \mathcal{X}.$$

- This is the foundation of our stationarity condition, which leads to the information-directed selection, i.e., **the second design principle**.
- This is known as *homogeneity of degree 1* or *constant returns to scale* in economics.

Summary and perspectives

Generalizations and applications: These two ingredients holds in great generality.

- In this talk, we discussed different **exploration queries**.
  - Multi-task BAI, e.g., with risk constraint[1].
- We can also extend to more general **reward feedback structure** (structured bandits):
  - Linear bandits and contextual bandits.
  - Markov chains, e.g., M/M/1 queue.
  - Dueling bandits with preferential feedback[2][3].
- We can also extend to accomodate more general **noise distributions**:
  - Single-parameter exponential family distributions and Heavy-tailed distributions[4].

---

[1] M. Hu and J. Hu, "Multi-Task Best Arm Identification with Risk Constraint," 2024.

[2] V. Dwaracherla, S. M. Asghari, B. Hao, and B. Van Roy, "Efficient exploration for LLMs," *arXiv preprint arXiv:2402.00396*, 2024.

[3] J. Liu, D. Ge, and R. Zhu, "Reward learning from preference with ties," *arXiv preprint arXiv:2410.05328*, 2024.

[4] S. Agrawal, S. Juneja, and P. Glynn, "Optimal \delta-Correct Best-Arm Selection for Heavy-Tailed Distributions", in *Algorithmic Learning Theory*, 2020, pp. 61–110.

Summary and perspectives

Methodology: Note that the original maximin optimization problem is *non-smooth*. However, we can derive an equivalent smooth optimization problem by introducing the dual variable $\boldsymbol{\mu}$ and reformulating the problem as

$$\Gamma_{\boldsymbol{\theta}}^{*} = \max_{\boldsymbol{p}\in\mathcal{S}_K} \min_{x\in\mathcal{X}} C_x(\boldsymbol{p};\boldsymbol{\theta}) = \max_{\boldsymbol{p}\in\mathcal{S}_K} \min_{\mu\in\mathcal{S}_{|\mathcal{X}|}} \sum_{x\in\mathcal{X}} \mu_x C_x(\boldsymbol{p};\boldsymbol{\theta}).$$

- This is a maximin concave-convex programming problem.
- It has rich connection with evolutionary game theory.
  - Indeed, we can view our problem as a two-play zero-sum game.
  - When $C_x(\boldsymbol{p};\boldsymbol{\theta})$ is linear in $\boldsymbol{p}$, it is called a bilinear game.
- We can formulate a continuous time version of an algorithm, whose evolution is governed by an ordinary differential equation or differential inclusion. The powerful tool of Lyapunov can be used to analyze convergence.

Summary and perspectives

Beyond pure exploration: Cost-aware exploration that bridges pure exploratoin and reward maximization[1].

- Consider assigning treatment to a large population of n individuals.
- Consider both **within-experiment cost** $C_i(\boldsymbol{\theta})$ and **post-experiment cost** $\Delta_i(\boldsymbol{\theta})$:

$$\mathrm{Cost}_{\boldsymbol{\theta}}(n, \pi) = \sum_{t=0}^{\tau-1} C_{I_t}(\boldsymbol{\theta}) + (n - \tau)\Delta_{\hat{I}_\tau}(\boldsymbol{\theta}).$$

- *BAI*: if $C_i = c$ and $\Delta_i = \theta_{I^*} - \theta_i$.
- *Reward minimization*: if $C_i = \varepsilon + (\theta_{I^*} - \theta_i)$ and let $\varepsilon \to 0$.
- As $n \to \infty$, we have $\mathrm{Cost}_{\boldsymbol{\theta}}(n, \pi) \sim \log(n) \cdot (\Gamma^*_{\boldsymbol{\theta}})^{-1}$, where $\Gamma^*_{\boldsymbol{\theta}}$ is again characterized by a similar maximin optimization problem.

---

[1] C. Qin and D. Russo, "Optimizing Adaptive Experiments: A Unified Approach to Regret Minimization and Best-Arm Identification." [Online]. Available: https://arxiv.org/abs/2402.10592

Thank you!

Access our full paper here: